STAT 300 Chapter 12

Zepu Zhang
April 12, 2012

1 Linear probabilistic model

X: predictor, independent/explanatory variable
Y: response, dependent variable

We consider X to be under our control; it can take various values but it is not random. To reflect this view, we use x when the predictor corresponding to a particular Y is needed, and use X mostly as the name of the predictor variable only.

Deterministic linear relationship: $y = \beta_0 + \beta_1 x$

Non-deterministic (probabilistic) linear relationship: see Ex. 12.1 and Ex. 12.2 in Chap. 12.1.

We model the response Y by the following assumed relationship between Y and X:

$$Y = \beta_0 + \beta_1 x + \epsilon, \quad \text{where } \epsilon \sim N(0, \sigma^2) \tag{1}$$

We can understand this model from several perspectives.

1. The model says that Y is a linear function of x plus a random fluctuation, ϵ. Moreover, the model assumes the fluctuation is normal, centered at 0, with variance σ^2 that does not change with the value of X.

2. From the model (1) we see

   $$E(Y) = E(\beta_0+\beta_1 x+\epsilon) = \beta_0+\beta_1 x+E(\epsilon) = \beta_0+\beta_1 x$$

   and

   $$\text{var}(Y) = \text{var}(\beta_0 + \beta_1 x + \epsilon) = \text{var}(\epsilon) = \sigma^2$$

   In addition, because ϵ is a normal random variable, so is Y. Hence, the model (1) is equivalent to

   $$Y \sim N(\beta_0 + \beta_1 x,\ \sigma^2) \tag{2}$$

After all, the focus of our concern is Y rather than ϵ, which is a device created in order to study Y. In general, the center of a statistical model is an assumption about the distribution of the variable of interest (here, Y).

The model (1) is a probabilistic one: given a particular value of X, we can't determine Y—Y is still uncertain. But we have an assumption about the statistical behavior (i.e. distribution) of Y conditional on the particular value of X, as in (2).

See Figure 12.4 in Chap. 12.1.

3. The Ys are independent of each other, because the ϵs are independent.

4. The expected value of Y has a deterministic, linear relation with X: $E(Y) = \beta_0 + \beta_1 x$. The model (1) is a linear regression model. Because there is only one X variable, it's called simple linear regression.

   The line
   $$y = \beta_0 + \beta_1 x$$
   is called the "regression function".

5. Interpretation of β_0 (intercept) and β_1 (slope): see p. 451 (7th ed) or p. 473 (8th ed). Usually we care more about β_1 than about β_0.

If we know the model parameters (β_0, β_1, and σ^2), we have a complete description of the behavior of the random variable Y, hence we can make certain probabilistic statements about Y, e.g. what is $P(Y > 3 \,|\, x = 1.3)$?

Example Ex. 12.3 in Chap. 12.1.

2 Estimating the regression coefficients

Suppose we have observed a bunch of y's at corresponding x values. We assume each y is a random "realization" (or "sample") of the conditional random variable $Y \,|\, x$, whose distribution is $N(\beta_0 + \beta_1 x,\, \sigma^2)$.

Note We are assuming a "data-generating" mechanism. By this assumption, corresponding to any fixed X level, say x, Y is a normal variable $N(\beta_0 + \beta_1 x, \sigma^2)$; our observed value y is a random sample from this distribution. (This also reinforces the interpretation about X: it's not a random variable; it's whatever level of the predictor that we "use".)

However, we don't know the true model, i.e., the parameters β_0, β_1, σ^2. We need to estimate them based on the data, (x_1, y_1),..., (x_n, y_n). What would be a sensible estimate?

The criterion: least squares. See first box in Chap. 12.2.

> Least squares estimation is an optimization problem. Take derivative, set to zero, and solve the so-called "normal equations" for the optimal values. See after box 1 and through box 2 in Chap. 12.2.

We may use the symbols $\hat{\beta}_0$ and $\hat{\beta}_1$ for the estimators (the general formulas) and b_0, b_1 for the estimates (the particular values after plugging in the actual data).

The LS estimators are (see (12.2)–(12.3) in Chap. 12.2.):

$$\begin{aligned} \hat{\beta}_1 &= \frac{\sum(x_i - \overline{x})(Y_i - \overline{Y})}{\sum(x_i - \overline{x})^2} = \frac{S_{xy}}{S_{xx}} \\ \hat{\beta}_0 &= \overline{Y} - \hat{\beta}_1 \overline{x} \end{aligned} \tag{3}$$

The estimates may be written as

$$\begin{aligned} b_1 &= \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sum(x_i - \overline{x})^2} = \frac{S_{xy}}{S_{xx}} \\ b_0 &= \overline{y} - b_1 \overline{x} \end{aligned}$$

Note (1) Get b_1 first, then b_0.
(2) The formula for b_0 suggests that the estimated regression line goes through the "average point" $(\overline{x}, \overline{y})$.
(3) Remember the definitions for the symbols S_{xx} and S_{xy}.

Cautions (1) Always make a scatter plot and get an intuitive sense of the relation.
(2) Danger of extrapolation (p. 458, 7th ed, or p. 480, 8th ed).

Example Ex. 12.4 in Chap. 12.2.

Example Ex. 12.5 in Chap. 12.2.

3 Estimating the variance

Following estimation of the regression coefficients, we have the <u>fitted values</u> $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ and <u>residuals</u> (or "errors") $e_i = y_i - \hat{y}_i$ (could be written as $\hat{\epsilon}_i$).

Note The residual is pointing from the fitted value (which lies on the fitted line) to the observed value. It is the deviation of y from its (estimated) expected value.

Example Ex. 12.6 in Chap. 12.2.

The variance σ^2 in the model (1) is estimated by

$$\hat{\sigma^2} = S^2 = \frac{\sum e_i^2}{n-2} \tag{4}$$

S^2 (or s^2) is called the MSE (mean squared error).

Why dividing by $n-2$? See top of p. 461 (7th ed) or middle of p. 483 (8th ed).

4 ANOVA for the simple regression model

Fact The LS regression function guarantees that $\sum e_i = 0$. Equivalently, $\sum \hat{y}_i = \sum y_i$ and $\overline{\hat{y}} = \overline{y}$.

Define

$$\begin{aligned}
\mathrm{SST} &= \sum (y_i - \overline{y})^2 \\
\mathrm{SSE} &= \sum (e_i - \overline{e})^2 = \sum e_i^2 \\
\mathrm{SSR} &= \sum (\hat{y}_i - \overline{\hat{y}})^2 = \sum (\hat{y}_i - \overline{y})^2
\end{aligned}$$

These quantities measure variations. Specifically, SST (total sum of squares) measures the total variation in the data. SSE (error sum of squares) measures variation in the data that is not "explained" by the regression model. SSR (regression sum of squares) measures variation "explained" by the linear relationship, i.e., variation in the fitted (or regressed, or predicted) Ys.

Note $s^2 = \frac{\mathrm{SSE}}{n-2}$

A fundamental relation:

$$\mathrm{SST} = \mathrm{SSE} + \mathrm{SSR}$$

This is a decomposition of the total variation into two parts, one part is explained by the model and the other is unexplained by the model.

Exercise Can you prove this relation?

Example Ex. 12.7 in Chap. 12.2.

Example Ex. 12.8 in Chap. 12.2.

Coefficient of determination

$$R^2 = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}}$$

Interpretation: see definition box (and after the box) on p. 463 (7th ed) or p. 485 (8th ed).

Example Ex. 12.9 in Chap. 12.2.

5 Some useful relations in computation

We first get data summaries n, $\sum x_i$, $\sum y_i$, $\overline{x}$, $\overline{y}$, $\sum x_i y_i$, $\sum x_i^2$, $\sum y_i^2$, then compute the following quantities, which is useful for various things. Below we list some useful or interesting relations.

$$
\begin{aligned}
S_{xy} &= \sum (x_i - \overline{x})(y_i - \overline{y}) \\
&= \sum (x_i y_i - x_i \overline{y} - \overline{x} y_i + \overline{x}\,\overline{y}) \\
&= \sum x_i y_i - n\overline{x}\,\overline{y} - n\overline{x}\,\overline{y} + n\overline{x}\,\overline{y} \\
&= \sum x_i y_i - n\overline{x}\,\overline{y} \\
S_{xx} &= \sum x_i^2 - n(\overline{x})^2 \\
\hat{\beta}_1 &= \frac{S_{xy}}{S_{xx}} \\
\hat{\beta}_0 &= \overline{y} - \hat{\beta}_1 \overline{x} \\
e_i &= \hat{\beta}_0 + \hat{\beta}_1 x_i - y_i \\
&= \overline{y} - \hat{\beta}_1 \overline{x} + \hat{\beta}_1 x_i - y_i \\
&= \hat{\beta}_1 (x_i - \overline{x}) - (y_i - \overline{y})
\end{aligned}
$$

(Hence $\sum e_i = \hat{\beta}_1 \sum (x_i - \overline{x}) - \sum (y_i - \overline{y}) = 0$.)

$$
\begin{aligned}
\text{SSE} &= \sum e_i^2 \\
&= \hat{\beta}_1^2 \sum (x_i - \overline{x})^2 + \sum (y_i - \overline{y})^2 - 2\hat{\beta}_1 \sum (x_i - \overline{x})(y_i - \overline{y}) \\
&= \hat{\beta}_1^2 S_{xx} + S_{yy} - 2\hat{\beta}_1 S_{xy} \\
&= S_{yy} - \hat{\beta}_1 S_{xy} \\
&= \sum y_i^2 - n\big(\overline{y}\big)^2 - \hat{\beta}_1 \Big(\sum x_i y_i - n\overline{x}\,\overline{y}\Big) \\
&= \sum y_i^2 - \hat{\beta}_1 \sum x_i y_i - n\overline{y}\big(\overline{y} - \hat{\beta}_1 \overline{x}\big) \\
&= \sum y_i^2 - \hat{\beta}_1 \sum x_i y_i - \hat{\beta}_0 \sum y_i \\
\text{SST} &= \sum (y_i - \overline{y})^2 \\
&= S_{yy} \\
&= \sum y_i^2 - n\big(\overline{y}\big)^2 \\
\text{SSR} &= \text{SST} - \text{SSE}
\end{aligned}
$$

6 Inferences about the slope parameter

"Inferences" include estimation, confidence interval, and hypothesis tests. The ability to do CI and tests relies on knowing the sampling distribution of the point estimator.

Example Ex. 12.10 in Chap. 12.3.

The sampling distribution of $\hat{\beta}_1$ (see (3)) is normal. Why?
P. 469–470 (7th ed) or p. 491–492 (8th ed).

$\hat{\beta}_1$ is a linear function of the Ys:

$$\hat{\beta}_1 = \sum_i c_i Y_i, \quad \text{where } c_i = (x_i - \overline{x})/S_{xx}$$

Because X is not random, every term involving X only is just a fixed number. Because the Y_is are normal variables, $\hat{\beta}_1$, which is a linear combination of the Y_is, is also normal. Its expected value is

$$E(\hat{\beta}_1) = E\Big(\sum c_i Y_i\Big) = \sum c_i E(Y_i) = \sum \frac{x_i - \overline{x}}{S_{xx}}(\beta_0 + \beta_1 x_i) = \cdots = \beta_1$$

Therefore $\hat{\beta}_1$ is an <u>unbiased</u> estimator. Its variance is

$$\text{var}(\hat{\beta}_1) = \text{var}\Big(\sum c_i Y_i\Big) = \sum c_i^2 \,\text{var}(Y_i) = \sum \frac{(x_i - \overline{x})^2}{S_{xx}^2}\sigma^2 = \frac{\sigma^2}{S_{xx}}$$

(In this derivation we have used the independence between the Y_is.)

To sum up, the sampling distribution of $\hat{\beta}_1$ is

$$\hat{\beta}_1 \sim N\big(\beta_1,\, \sigma^2/S_{xx}\big)$$

Interpretation: $\hat{\beta}_1$ is a random variable as we re-sample Y from the conditional distributions $p(Y \mid x)$ (sticking with the set of fixed x values) and re-calculate the LS estimates.

Standardize:

$$\frac{\hat{\beta}_1 - \beta_1}{\sigma/\sqrt{S_x x}} \sim N(0, 1)$$

In all realistic situations, σ is unknown. Plug in its estimate, $s = \sqrt{s^2}$ as defined in (4). Then, like before, we have a t variable (p. 471, 7th ed.; p. 493, 8th ed.):

$$\frac{\hat{\beta}_1 - \beta_1}{s/\sqrt{S_{xx}}} \sim t_{n-2}$$

(The df of this t distribution is related to the $n - 2$ that appears in the definition of s^2.)

With the sampling distribution of $\hat{\beta}_1$ and the re-arrangements above, CI and tests for β_1 are similar to what we have learned in Chaps. 7.3 and 8.2. See formulas in boxes on p. 471 and p. 474 (7th ed) or p. 493–494 and p. 496.

Example Ex. 12.11 in Chap. 12.3.

Example Ex. 12.12 in Chap. 12.3.

7 Correlation

7.1 Sample correlation coefficient

Definition

$$r = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}}$$

For interpretations of the numerator, and the rational for the normalization by such denominator: see p. 485–486 (7th ed) or p. 508–509 (8th ed).

Example Ex. 12.15 in Chap. 12.5.

Properties of r and interpretations: in Chap. 12.5.

7.2 Population correlation coefficient

Definition

$$\rho = \frac{\mathrm{cov}(X, Y)}{\sigma_X \cdot \sigma_Y}$$

A usual estimator of ρ is the sample correlation coefficient R. (Write the estimator as R and the estimate as r.)

Example Ex. 12.16 in Chap. 12.5.

(Skip inferences about ρ.)

8 Useful `R` functions

`lm, coef, fitted, resid, abline, cor`

Read a short and nice tutorial at http://www.cyclismo.org/tutorial/R/linearLeastSquares.html.